Review Article

Whole Genome Sequence Analysis of Viruses; Moving Beyond Single/Partial Gene Based Phylogenies in Context of Epidemiology and Genetic Evolution

Sajad Ahmad Wani[1*], Amit Ranjan Sahu[1], Bishnu Prasad Mishra[1], Amod Kumar[2], Govindarajan Bhuvana Priya[3], Arpita Padhy[4], Aditya Prasad Sahoo[1], Ashok Kumar Tiwari[5], Ravi Kumar Gandham[1], Raj Kumar Singh[1]

[1]Division of Veterinary Biotechnology; [2]Division of Animal Genetics; [3]Division of Bacteriology and Mycology; [5]Division of Standardization, ICAR-Indian Veterinary Research Institute, Izatnagar, Bareilly, UP-243122; [4]Department of Veterinary Microbiology, Aarawali Veterinary College, Sikar, Rajasthan- 332001, India.

Abstract | The enormous amount of viral species in nature arouses curiosity about not only their origin, but also forces their naming and organizing them into hierarchically arranged systematic units. The rapid evolution of viruses, in particular RNA viruses, has led to emergence of many new genotypes. The sequencing of whole genomes, genes or gene fragments is more and more commonly used for understanding epidemiology. Most accepted modern phylogenies are derived using sequences from individual homologous genes which may fail sometimes to construct a true phylogenetic tree. The first problem is that the evolutionary history of a particular gene is not necessarily the same as the evolutionary history of the virus in which it can be observed. This might be due to duplication and deletion, or even horizontal gene transfer between different viruses. Secondly, it is not always possible to find genes that are sufficiently conserved across all our viruses of interest to be successfully identified, and yet sufficiently diverged to be of use for phylogenetic analysis. One of the most pervasive challenges in molecular phylogenetics is the incongruence and discordance between phylogenies obtained using different data sets, such as individual genes. Whole genome based phylogenetic analysis has helped to characterize the novel viruses, uncover the population history of the disease, elucidate virus-host interactions, understanding of the evolutionary rates, monitoring gene reassortment, interspecies transmission between different viral strains and produce estimates of epidemiological parameters. In this review, we discussed efforts that have been made to infer phylogenies by consideration of the viruses at the genome level, rather than just individual genes.

Keywords | Phylogeny, Epidemiology, Evolution, Reassortment, Genetic diversity

Editor | Kuldeep Dhama, Indian Veterinary Research Institute, Uttar Pradesh, India.
Received | June 05, 2015; **Revised** | June 22, 2015; **Accepted** | June 23, 2015; **Published** | June 30, 2015
***Correspondence** | Sajad Ahmad Wani, ICAR-Indian Veterinary Research Institute, Izatnagar, Bareilly, UP, India; **Email:** wanisajad759@gmail.com
Citation | Wani SA, Sahu AR, Mishra BP, Kumar A, Priya GB, Padhy A, Sahoo AP, Tiwari AK, Gandham RK, Singh RK (2015). Whole genome sequence analysis of viruses; moving beyond single/partial gene based phylogenies in context of epidemiology and genetic evolution. Adv. Anim. Vet. Sci. 3(8): 435-443.
DOI | http://dx.doi.org/10.14737/journal.aavs/2015/3.8.435.443
ISSN (Online) | 2307-8316; **ISSN (Print)** | 2309-3331

INTRODUCTION

Molecular phylogeny is a fundamental tool for understanding evolution, relationships and assigning of genotypes. Phylogenetic reconstructions have been used mainly to differentiate and classify isolates below the species level, mainly into genotypes. Despite being designed to study evolutionary relationships amongst taxa, molecular phylogenetic analysis also provides a framework to show not only sequence similarity but also the clustering and relationships amongst different isolates. In epidemiological studies, a stricter phylogenetic approach should be used to establish the evolution of different sequences over time. Until recently, the method commonly used for molecular phylogenetic studies for many viruses typically involved sequencing a gene or partial genomic region (phylogenetic marker) and inferring a "gene tree" based on these sequences and declaring the gene tree to be the estimate of the tree of strain relationships (Degnan et al., 2009; Luo et al., 2009; Padhi et al., 2009). Phylogeny in-

ferred from a gene or protein sequence only describes the evolution of that particular gene or encoded protein. This sequence may evolve more or less rapidly than other genes in the genome or may have a different evolutionary history from the rest of the genome owing to horizontal gene transfer events. Single gene phylogenies were used because of incomplete genome sequence information and the inherent limitations of available computer programs; however, many studies have shown that the evolutionary histories of some individual genes or genomic regions may not be identical to each other within many viruses, which may be due to the recombination, reassortment and selection pressure (Herniou et al., 2001; Magiorkinis et al., 2004; Olvera et al., 2007; Olvera et al., 2010; Anderson et al., 2010; Tatte et al., 2010). Earlier due to high cost of sequencing and less efficient platforms the complete genome sequencing was not possible for every virus. However the decrease in cost and advent of high throughput sequencing led blast of whole genomes in the public data bases. It is full of ambiguity for phylogenetic analysis based on single gene when using conserved or similar genes since horizontal gene transfer (HGT) between viruses, along with gene duplication, gene capture from host appears to have been frequent in large DNA viruses (Herniou et al., 2001; Filee et al., 2003; Shackelton et al., 2004). It is important to consider the possibility of genetic recombination while evaluating apparent phylogenetic relationships between viral strains. Complete genome sequences contain phylogenetic information at several levels. In addition to the nucleotide sequence and amino acid sequences of the encoded proteins, the gene content and the order of genes on a genome may be phylogenetically informative (Koonin et al., 2000; Rokas et al., 2000). Gene content or gene order data sets are independent of these sequences of individual genes and should complement phylogenies based on nucleotide or amino acid sequences. Several attempts have been made to infer viral phylogeny from their whole genomes to avoid the problem of gene rearrangement, gene loss, gene duplication and lateral gene transfer. However, some of them infer the majority consensus tree of the many trees of individual genes or use the combined sequences of many shared genes. Some of them employ gene content and gene order method, but the former has to correct for the genome size effect and the latter can be hindered by a lack of synteny conservation or the variation of the evolving rate of synteny between taxa (Montague et al., 2000; Gao et al., 2003; Harrison et al., 2003; Snel et al., 2005). Therefore, several methods such as phylogenetic networks have been developed to infer these evolutionary processes including the recombination events. However, all these analyses could only provide accurate estimates only when all the whole-genome sequences are available (Olvera et al., 2010). Molecular epidemiology using whole genome sequences of pathogens will reveal more precise phylogenetic relationships as compared to gene or partial sequences, thus giving an exact picture of geographical and evolutionary origin of the viral isolates. Phylogenetic analysis is a prerequisite for virus tracing and thus allows implementing more effective control measures.

NEXT GENERATION SEQUENCING

With the advent of next generation sequencing technologies and falling costs of sequencing, a paradigm shift has taken place from traditional Sanger's method to whole genome sequencing. Compared to the traditional Sanger capillary sequencer, next-generation sequencers are capable of massively parallel sequencing of millions of amplified DNA molecules in a single run and they do not require the conventional cloning and amplification. Next generation sequencing is currently driven by 454 GS FLX titanium (Roche), Genome Analyzer-II (Illumina/Solexa), ABI-SOLiD (Life technologies-Applied Biosystems), Polonator G 007 (Danaher motions), Heliscope (Helicose Bioscience) and Pac Bio RS(Pacific Biosciences, SMRT-single molecule Real time sequencing technology) (Pacific Biosciences) and Ion Torrent (Life Technologies) (Table 1). Despite their different configurations, the next-generation sequencers share many common features: (1) relatively small amount of starting DNA (a few micrograms) is needed, (2) fragmented DNA templates are ligated to specific adaptors at both ends, (3) multiple PCR amplification cycles are performed, (4) amplified DNA templates are attached to a solid support in a reaction chamber or a flow cell, (5) during the extension cycles, sequencing reagents are repetitively applied and washed away and (6) the number of the extension cycles is often limited, thus producing shorter read lengths of 35-250 bases as compared to the read length of 650-800 bases in the Sanger capillary sequencing. Because of their high-throughput capacities, these next-generation sequencers are better suited for the studies dealing with the whole genomes, replacing the Sanger sequencing in many situations.

NGS PLATFORMS

Most commonly sequencing platforms nowadays used are as: Roche 454 FLX Pyrosequencer, Illumina sequence, ABI SOLiD System. The Roche 454 sequencing technology combines the principles of emulsion PCR and pyrosequencing. The steps involve fragmentation of the template DNA, ligation to adaptors and clonal amplification of DNA using emulsion PCR. The emulsion beads are then deposited in picotiter plate wells containing smaller beads with sequencing enzyme and buffers required to perform iterative pyrosequencing – translating each nucleotide incorporation event into a well-specific pyrophosphate-tagged luminescence. The newer and robust 'titanium' chemistry can generate 1×10^6 sequence reads of longer read length (≥400 bp), yielding 500 million bp of sequence per run.

Table 1: Comparison of the Next-Generation DNA sequencing platforms

Platforms	Starting DNA (µg)	Amplification	Sequencing method	Read length (bases)	Throughput capability (Gb per run)	Run time (days)
Roche (454) GS-FLX	3-5	Emulsion PCR	Pyrosequencing	330	0.45	0.35
Illumina Genome Analyzer II system	0.1-1	Bridge PCR	Sequencing by synthesis (Reversible termination)	75-100	18-35	4^{F} ,9^{M}
ABI SOLiD	0.1-20	Emulsion PCR	Sequencing by ligation	80	30-50	7^{F},14^{M}
Polonator G007	15*	In situ slide PCR/ Emulsion-PCR-beads/ Isothermal amplification	Sequencing by ligation	26	12	5^{M}
Helicos Biosciences (Heliscope)	1-5	Single molecule	Sequencing by synthesis (Reversible termination)	32	37	8^{F}
SMRT(Pacific Biosciences)	2-3	Single Molecule	Real time	964		
Ion Torrent	0.1	NA	Semiconductor Real time sequencing	200	1	0.125

NA – Not available; **F** – Fragment library; **M**– Mate pair Library

The Illumina/Solexa Genome Analyzer was the first 'short read' sequencing platform commercially available that involves sequencing-by-synthesis using reversible terminators. Fragmented ssDNA is hybridized to oligonucleotide anchors on a solid surface referred to as a 'flow cell'. Solid-phase bridge amplification of DNA templates is conducted to generate amplified clusters. Massively parallel sequencing of cleaved products from amplified clusters is carried out using DNA polymerase and a set of four base-specific color-coded reversible terminators that result in growing oligonucleotide chains. This platform originally produced 35-bp reads to yield 1 Gb of sequence output per 2–3-day run. Subsequent upgrades on this platform have increased both the density of clusters and read lengths so that this machine can currently yield 4 Gb of sequence output in a 2–3-day run.

The third commercially available platform is the ABI SOLiD platform, which uses hybridization–ligation methodologies for massively parallel sequencing. The initial emulsion PCR step is the same as that in the Roche 454 platform, except that the beads are only 1 µm in size. The amplified product on the beads is then covalently linked to a glass surface, and sequencing is carried out using hybridization–ligation with an octamer interrogation probe consisting of two probe-specific, three degenerate and three promiscuous bases. Each nucleoide position is ascertained using a four-dye encoding schema and each position is interrogated twice to distinguish sequencing errors from single nucleotide polymorphisms (SNPs). When first available (early 2007), this platform had an output in short reads of 35 bp and produced 1–3 Gb of sequence data per 8-day run. The current upgrade on this system (now the SOLiD 3) is capable of a much higher density of beads and has an output of 20–40 Gb per 8–10-day sequence run.

NGS DATA ANALYSIS

NGS experiments generate unprecedented volumes of data, which present challenges and opportunities for data management, storage, and most importantly, analysis. Data volumes generated during single runs of the 454 GS FLX, Illumina, and SOLiD instruments are approximately 15 GB, 1 TB, and 15 TB, respectively a large variety of software programs for alignment and assembly have been developed and made available to the research community. Most use the Linux operating system, and a few are available for Windows. Many require a 64-bit operating system and can use 16GB of RAM and multiple central-processing unit cores. The range of data volumes, hardware, software packages, and settings leads to processing times from a few minutes to multiple hours, emphasizing the need for sufficient computational power. Although a growing set of variations in alignment and assembly algorithms are available, there remains the trade-off between speed and accuracy in which many but not all possible alignments are evaluated, with a balance having to be struck between ideal alignment and computational efficiency.

Softwares and Bioinformatics Tools for Data Analysis

A variety of software tools are available for analyzing next-generation sequencing data like Cross_match, ELAND, Exonerate, MAQ, Mosaik, RMAP, SHRiMP, SOAP, SSAHA2, SXOligoSearch, Synamatix, ALLPATHS, Euler-SR, SHARCGS, SHRAP, SSAKE, VCAKE, Bowtie, Velvet, PyroBayes, PbShort etc. Their functions fit into several general categories, including: (i) alignment of sequence reads to a reference; (ii) base-calling and/or polymorphism detection; (iii) *de novo* assembly, from paired or

unpaired reads; and (iv) genome browsing and annotation.

Technical Problems and Limitations of NGS

There are some common technical problems associated with various NGS platforms. Short reads in many NGS systems result in difficulties with assembling and mapping to the reference sequences, particularly at repetitive regions. Not all sequences are equally processed and sequenced, and DNA regions enriched with GC content are particularly prone to low coverage. For NGS platforms with target amplification or enrichment, amplification bias may be introduced. Last but not least, sequencing errors are present essentially in all NGS platforms. Longer reads are prone to have error readings, particularly towards the ends. Repetitive sequences and homopolymers are also of concern for some third generation sequencers; however, rapid improvement has been made to overcome these problems. Increase of coverage and deep sequencing are important to correct some of these problems.

WHOLE GENOME SEQUENCING OF VIRUSES OF VETERINARY IMPORTANCE

In recent years several pathogens of veterinary importance have been sequenced world over. Whole-genome sequencing of microbes has revolutionized the methods by which these organisms are studied and has heightened expectations regarding the ability to predict potential targets for antimicrobial agents and vaccines. It is now possible to sequence viral genomes or sample entire transcriptomes more efficiently and in greater depth than ever before. It is less expensive, quicker and more efficient to access gene sequences by whole genome sequencing than traditional gene-by-gene approaches. Molecular epidemiology using whole genome sequences of pathogens will reveal more precise phylogenetic relationships as compared to gene or partial sequences, thus giving an exact picture of geographical and evolutionary origin of the viral isolates. The number of complete genomes of viral/bacterial pathogens has increased dramatically in recent years with submission of enormous sequence data in sequence repositories such as GenBank. Currently 2703 viral whole genome sequences available in the public domain at NCBI e.g. Foot and mouth disease virus, Peste des petits ruminants virus, Rinderpest virus, Rabies virus, Canine distemper virus, Sheep pox virus, Goat pox virus , Bluetongue virus, New castle disease virus, Infectious bursal disease virus, Classical swine fever virus, Bovine viral diarrhea virus.

METHODS OF PHYLOGENETIC TREE CONSTRUCTION

Neighbour-joining, Maximum likelihood and Bayesian approach are most commonly used for constructing phylogenetic trees. The accuracy of the tree-building methods used for phylogenetic analysis depends on the assumption on which each the method is based. Understanding these assumptions is the first step toward efficient use of these methods. The second step is understanding, how the methods actually work and what intrinsic limitations these methods have. The third step is choosing suitable phylogenetic method(s) that can give a reasonably correct picture of a phylogenetic tree. Neighbour-joining is one of the distance-based methods. It is extremely fast and has been advocated for analysis of large datasets. However, recovery of the true tree is guaranteed only if the distance matrix is correct, and calculation of genetic distances is complicated by biological processes such as rate heterogeneity. It is thus not recommended for use in finding final tree. Maximum likelihood (ML), one of the character-based methods, has also been used for phylogenetic analysis of pestiviruses. Under an evolutionary model, the most probable tree is found by an optimality criterion based on the character (nucleotide) at each position of a set of sequences. Disadvantages using ML are that it is computationally intensive when dealing with many taxa, and may yield unreliable results with regard to complex parameter-rich model robustness of the so-called best tree can be estimated statistically by bootstrapping (e.g. 1000 replicates) the original dataset and a value of more than 70% is thought to indicate support for a group on the tree. The Bayesian approach has been recently developed for inferring phylogeny. It is rapidly accepted in phylogenetics. In contrast to the traditional ML method that only gives the topology of a tree, the Bayesian analysis produces both a tree estimate and a measurement of uncertainty for the groups on the tree, thus providing a measure of support faster than ML bootstrapping. By using a Markov chain Monte Carlo (MCMC) algorithm, the Bayesian phylogenetic inference allows implementation of complex parameter-rich evolution models. It is important to realize that phylogenetic tree reconstruction is not a trivial matter, but a complicated process that often requires careful thought. Accuracy, reliability, and computational speed are all major factors for consideration when choosing a particular phylogenetic method. It is also important to realize that none of the three phylogenetic reconstruction methods are guaranteed to find the correct tree. All three methods have the potential to produce erroneous trees. To minimize phylogenetic errors, it is recommended that at least two methods be used for any phylogenetic analysis to check the consistency of tree building results obtained.

WHOLE GENOME BASED APPROACH FOR PHYLOGENETIC ANALYSIS

Complete genome approaches have recently been employed to infer the phylogeny of many viruses. Wang et

al. (2013) identified discordance between full-length genome tree and individual gene trees upon phylogenetic analysis of hepatitis E virus (HEV), Japanese encephalitis virus (JEV), measles virus (MV) and porcine circovirus 2 (PCV2). For all the four viruses the individual gene trees differed not only from the corresponding genome tree, but from the trees constructed from the other genes in the same genome, in both their topologies and branch lengths in a way. In HEV, the trees of region GO and KLY-B differed dramatically from the genome trees in topologies, which resulted in misleading inferences on genetic relationships of some strains. However, it was hard to estimate which of the regions SGG-A, MJ-C and MXJ produced a tree that was most similar to the full genome tree. In JEV, the trees of gene E, NS1 and NS5 could agree well with the genome tree. However, it was hard to access which one from the genes NS2a, NS2b, NS3, NS4b and PreM could yield a tree more concordant with the genome tree due to different discordance involving different virus strains. The cap gene tree displayed topology obviously disagreeing with the genome tree and other gene trees. The MV trees based on the P gene, M gene, V gene and C gene could not match the genome trees very well. However, the L gene, N gene, H gene and F gene trees could reproduce the topology of the full-length genome tree more similar than others with reliable bootstrap support values. For PCV2, the cap gene shared a more similar tree with the genome than the rep gene both in topology and branch length obviously. However, the tree based on the rep gene presented much incongruence with the cap and the complete genome trees, leading to drastically disordered groupings for viral strains.

Historically the genotyping of Polyomavirus BK is carried out by VP1 single gene based phylogenies. Luo et al. (2009) revaluated the phylogenetic analysis using specific genes and BKV whole-genome of 162 sequences available in the public domain and found that BKV subtypes and subgroups can no longer be reliably determined by sequencing certain partial gene sequences. Based on whole genome based phylogenies Polyomavirus BK viral strains were clustered into previously defined subtypes I, II, III, and IV with high bootstrap values. Subtypes V and VI, previously classified on the basis of more limited sequence data, are best considered to be subgroups of subtype I and clustered with Ib2 and Ib1, respectively. Lower bootstrap separation values were obtained for subtype II (60% for VP1 versus 100% for the whole genome), Ib (55% for VP1 versus 87% for the whole genome), IVa (34% versus 83%), and IVc (21% versus 68%). Subgroups IVb1 and IVb2 clustered together in the whole-genome tree, IVb2 strains separated out in the VP1 trees and clustered together with IVc strains. Thus, to completely define these subtypes and subgroups, genetic information extending beyond the VP1 region is needed. Thus, with a rapidly expanding database of DNA sequences, a genotyping schema based entirely on VP1 can no longer adequately capture the genetic diversity of BKV. VP2 gene region sequences could resolve the subtypes but not the subgroups. The VP3 gene is located within VP2 gene and contains most of its informative sites. Phylogenetic trees based on small-T-antigen sequences can divide BKV into major subtypes and all subgroups of subtype I but cannot resolve subgroups of subtype IV.

A comprehensive phylogenetic analysis of 22 complete JCV genomes was accomplished first time by Jobes et al. (1998) using neighbour-joining, UPGMA and maximum parsimony methods. European Type 1 strains was found to be diverged from other subtypes during evolution .Previously phylogeny was carried out by most variable small V±T intergenic region (610 bp) nevertheless showed little variability between closely related JCV strains and may not provide enough informative sites (Sugimoto et al., 1997). Utilizing the whole JCV genome, minus the regulatory region (4854 bp), substantially increases the number of phylogenetically informative sites and more adequately resolves relationships between the JCV genotypes. Parsimony analysis showed that of the 611 total characters (a single gap was required in the alignment), 534 sites were invariant, 36 were phylogenetically uninformative and only 41 sites were informative. In contrast, of the 4856 characters in the whole genome data set, 4523 were invariant between the strains, 161 were uninformative and 172 sites were phylogenetically informative. The whole genome approach, therefore, provides a fourfold increase of informative sites over the V±T region alone. This increase in informative sites translates into a much better resolved phylogeny for JCV. V±T region sequences placed strain Tai-3 with the Type 3 group and assigned Type 2 strain g224A an ambiguous and unresolved position in the UPGMA and neighbour-joining trees.

Most analysis of baculovirus phylogeny has been based on the *polyhedrin/granulin* gene but other genes have been used (Bulach et al., 1999; Bideshi et al., 2000). Comparison of these analyses reveals that conflicts are often observed between phylogenies based on different genes. These conflicts could be due to erroneous phylogenetic inferences caused by unequal rates of evolution, lack of an unambiguous phylogenetic signal in the sequences or due to recombination. Exchange of genetic material is known to occur between coinfecting baculo viruses or between baculo viruses and their hosts (Fraser et al., 1995). A key question is to what extent gene exchanges have shaped the phylogeny of virus and is it possible to construct a single phylogenetic tree representing their evolutionary history. Despite the fluidity of bacula virus genome, the whole genome based methods are sufficiently powerful to unravel the underlying phylogeny of the viruses. Herniou et al. (2001) highlighted the fluid nature of baculovirus genomes, with evidence of frequent genome rearrangements and multiple gene content

changes during their evolution by whole genome based phylogeny.

The complete genome sequencing of the reference strain of bluetongue virus (BTV) serotype 16 (strain RSArrrr/16) was carried out by Maan et al. (2012). Previous phylogenetic comparisons show that BTV RNA sequences cluster according to the geographic origins of the virus isolate/lineage, identifying distinct BTV topotypes. Sequence comparisons of segments Seg-1 to Seg-10 show that RSArrrr/16 belongs to the major eastern topotype of BTV (BTV-16e) and can be regarded as a reference strain of BTV-16e for phylogenetic and molecular epidemiology studies. All 10 genome segments of RSArrrr/16 group closely with the vaccine strain of BTV-16 (RSAvvvv/16) that was derived from it, as well as those recently published for a Chinese isolate (Yang et al., 2011) of BTV-16 (>99% nucleotide identity), suggesting a very recent common ancestry for all three viruses.

The evolutionary dynamics of influenza A virus are shaped by a complex interplay between rapid mutation, frequent reassortment, widespread gene flow, natural selection (occasionally generating genome-wide selective sweeps), functional interactions among segments, and global epidemiological dynamics. Large scale phylogenetic analysis based on whole genome play a pivotal role in understanding the reassortment and evolution of influenza A virus. The co-existence and circulation of different lineages is a big hurdle in understanding the epidemiology of the virus. Whole genomes of H3N2 influenza A viruses sampled during 1999–2004 has identified two key evolutionary patterns (Holmes et al., 2005). Whole-genome analysis of human influenza A virus revealed multiple persistent lineages and reassortment among recent H3N2 Viruses. First, although the majority of viruses isolated after 2002 fall into a single phylogenetic group (clade A), multiple, co-circulating viral lineages are present at particular time points. The genetic diversity of influenza A virus is therefore not as restricted as previously suggested, particularly when genes other than that encoding HA are analysed. This co-circulation of lineages is most apparent with the identification of three clades of H3N2 viruses that appear to infect the same populations until 2002, after which they acquired a common HA gene through reassortment. Second, and more dramatically, these multiple, co-circulating lineages may have complex genealogical histories and interact through reassortment. Two reassortment events involving the HA gene of clade B: one in which it was acquired by the clade A viruses and another in which it was independently acquired by those isolates assigned to clade C. The utility of whole-genome analyses of influenza A viruses, and further makes clear that additional whole-genome analyses are required to understand fully the evolutionary mechanisms and epidemiological dynamics of this virus. While antigenic variance of HA is still the dominant selective pressure on human influenza A virus evolution, the finding that antigenically novel clades emerge by reassortment among persistent viral lineages rather than via antigenic drift is of major significance for vaccine strain selection.

Studies on genetic diversity of rotaviruses have been primarily based on the genes encoding the antigenically significant VP7 and VP4 proteins. Since the rotavirus genome has 11 segments of RNA that are vulnerable to reassortment events, analyses of the *VP7* and *VP4* genes may not be sufficient to obtain conclusive data on the overall genetic diversity, or true origin of strains. In the last few years following the advent of the whole-genome-based genotype classification system, the whole genomes of at least 167 human group A rotavirus strains have been analysed, providing a plethora of new and important information on the complex origin of strains, inter- and intra-genogroup reassortment events, animal– human reassortment events, zoonosis, and genetic linkages involving different group A rotavirus gene segments (Ghosh et al., 2011; Wang et al., 2014; Thongprachum et al., 2013; Matthijnssens et al., 2008). Recently Wang et al. (2014) carried out first large-scale whole genome-based study to assess the long-term evolution of common human rotaviruses (G3P[8]) in an Asian country from 2000 through 2013 and concluded Chinese G3P[8] rotavirus strains have evolved since 2000 by intra-genogroup reassortment with co-circulating strains, accumulating more reassorted genes over the years. The genetic information in this study is expected to contribute as a baseline data to understand long-term evolution of rotavirus genome and to formulate policies for the use of rotavirus vaccines. Studies on genetic diversity of rotaviruses have been primarily based on the genes encoding the antigenically significant VP7 and VP4 proteins. Since the rotavirus genome has 11 segments of RNA that are vulnerable to reassortment events, analyses of the VP7 and VP4 genes may not be sufficient to obtain conclusive data on the overall genetic diversity, or true origin of strains. In the last few years following the advent of the whole-genome-based genotype classification system, the whole genomes of at least 167 human group A rotavirus strains have been analysed, providing a plethora of new and important information on the complex origin of strains, inter- and intra-genogroup reassortment events, animal– human reassortment events, zoonosis, and genetic linkages involving different group A rotavirus gene segments (Ghosh et al., 2011; Wang et al., 2014; Thongprachum et al., 2013; Matthijnssens et al., 2008).

Recently Gao et al. (2015) elucidated the genetic diversity Japanese encephalitis virus by whole genome based phylogenetic analysis using Bayesian Markov chain Monte Carlo simulations. The results showed that the most recent common ancestor (TMRCA) for JEV was estimated to

have occurred 3255 years ago. Chronologically, this ancestral lineage diverged to produce five recognized virus genotypes in the sequence 5, 4, 3, 2 and 1. Population dynamics analysis indicated that the genetic diversity of the virus peaked during the following two periods: 1930–1960 and 1980–1990, and the population diversity of JEV remained relatively high after 2000. Genotype 5 is the earliest recognized JEV lineage, and the genetic diversity of JEV has remained high since 2000.

Detailed Bayesian coalescent phylogenetic analyses was performed on 97 whole-genome sequences to comprehensively examine molecular evolutionary rates and estimate dates of common ancestry for viruses within the family *Filoviridae* by Carroll et al. (2014). Molecular evolutionary rates for viruses belonging to different species range from 0.46×10^{4} nucleotide substitutions/site/year for *Sudan ebolavirus* to 8.21×10^{4} nucleotide substitutions/ site/year for *Reston ebolavirus*. Most recent common ancestry can be traced back only within the last 50 years for *Reston ebolavirus* and *Zaire ebolavirus* species and suggests that viruses within these species may have undergone recent genetic bottlenecks. Examination of the whole family suggests that members of the *Filoviridae*, including the recently described Lloviu virus, shared a most recent common ancestor approximately 10,000 years ago. These data will be valuable for understanding the evolution of filoviruses in the context of natural history as new reservoir hosts are identified and, further, for determining mechanisms of emergence, pathogenicity, and the ongoing threat to public health.

CONCLUSION, FUTURE ISSUES AND CHALLENGES AHEAD

The dynamic development of the whole genome phylogenetic analysis triggered a breakthrough in the perception of the world around us. Phylogenetic analysis based on single gene is full of ambiguity and may fail to fully reflect the current taxonomical classification of viruses. Comparison of these analyses reveals that conflicts are often observed between phylogenies based on different genes. These conflicts could be due to erroneous phylogenetic inferences caused by unequal rates of evolution or to lack of an unambiguous phylogenetic signal in the sequences. Single gene phylogenies reveal extensive incongruence and conflicting topologies. The availability of complete genome sequence data due to cheaper sequencing technologies for several viruses has led to an interest in the use of such data for phylogenetic reconstruction. Full-genome based analysis can provide the relatively more reliable information about genetic relationships between different virus isolates and determination of the directions of virus migrations from one country or continent to another. Whole-genome analysis is beneficial for molecular characterization and understanding of the evolution of the pathogen. It is also useful for monitoring gene reassortment and interspecies transmission between different viral strains. The genetic information in whole genome based phlogenetic analysis is expected to contribute as a baseline data to understand long-term evolution of viral genome, and to formulate policies for the use of vaccines. Phylogenetic inference using whole genome data poses tremendous statistical and computational challenges. There is a profound need to develop new models for the analysis of multigene or multipartition datasets that can accommodate factors such as the heterogeneity of the evolutionary process among genes, or partitions, in whole genome phylogenetic analyses. Improved statistical methods are needed that account for genomic variation in evolutionary rates, transition/transversion rate ratios, and local gene trees. Moreover, there is an urgent need to develop efficient computer programs for combined analysis of multipartition datasets, particularly those suitable for parallel computer systems. Hence whole genome based phylogenetic analysis is the need of hour to study the molecular epidemiology and genetic evolution of the animal viruses.

ACKNOWLEDGEMENT

The authors acknowledge the financial support provided by Department of Biotechnology (DBT), Government of India.

CONFLICT OF INTEREST

The authors have no conflict of Interest.

AUTHORS CONTRIBUTION

Bishnu Prasad Mishra, Aditya Prasad Sahoo, Ashok Kumar Tiwari, Ravi Kumar Gandham, Raj Kumar Singh gave the concept of the manuscript and overviewed the final manuscript. Sajad Ahmad Wani and Amit Ranjan Sahu drafted and revised the whole manuscript. Amod Kumar, Govindarajan Bhuvana Priya and Arpita Padhy equally contributed matter to the manuscript. All the authors have read and approved the manuscript.

REFERENCES

- Anderson CN, Liu L, Pearl D, Edwards SV (2012). Tangled trees: the challenge of inferring species trees from coalescent and noncoalescent genes. Methods Mol. Biol. 856: 3–28. http://dx.doi.org/10.1080/01431161.2010.542198
- Bideshi DK, Bigot Y, Federici BA (2000). Molecular characterization and phylogenetic analysis of the *Harrisina brillians* granulovirus granulin gene. Arch. Virol. 145(9): 1933–1945. http://dx.doi.org/10.1007/s007050070067
- Bulach DM, Kumar, CA, Zaia A, Liang B, Tribe DE (1999).

Group II nucleopolyhedrovirus subgroups revealed by phylogenetic analysis of polyhedron and DNA polymerase gene sequences. J. Invertebr. Pathol. 73(1): 59–73. http://dx.doi.org/10.1006/jipa.1998.4797

•Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, Swanepoel R, Rollin PE, Nichol ST (2013). Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J. Virol. 87(5): 2608–2616. http://dx.doi.org/10.1128/JVI.03118-12

•Degnan JH, Rosenberg NA (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24(6): 332–340. http://dx.doi.org/10.1016/j.tree.2009.01.009

•Filee J, Forterre P, Laurent J (2003). The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res. Microbiol. 154(4): 237-243. http://dx.doi.org/10.1016/S0923-2508(03)00066-4

•Fraser MJ, Cary L, Boonvisudhi K, Wang H (1995). Assay for movement of lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. Virology. 211(2): 397–407. http://dx.doi.org/10.1006/viro.1995.1422

•Gao L, Qi J, Wei H, Sun Y, Hao B (2003). Molecular phylogeny of coronaviruses including human SARS-CoV. Chinese Sci. Bull. 48: 1170-1174. http://dx.doi.org/10.1007/BF03183929

•Gao X, Liu H, Li M, Fu S, Liang G (2015). Insights into the evolutionary history of Japanese encephalitis virus (JEV) based on whole-genome sequences comprising the five genotypes. Virol. J. 12: 43. http://dx.doi.org/10.1186/s12985-015-0270-z

•Ghosh, Nobumichi K (2011). Whole-genomic analysis of rotavirus strains: current status and future prospects. Future Microbiol. 6(9): 1049-1065. http://dx.doi.org/10.2217/fmb.11.90

•Harrison RL, Bonning BC (2003). Comparative analysis of the genomes of Rachiplusiaou and Autographa californica multiple nucleopolyhedroviruses. J. Gen. Virol. 84(7): 1827-1842. http://dx.doi.org/10.1099/vir.0.19146-0

•Herniou EA, Luque T, Chen X, Vlak JM, Winstanley D, Cory JS, O'Reilly DR (2001). Use of whole genome sequence data to infer baculovirus phylogeny. J. Virol. 75(17): 8117-8126. http://dx.doi.org/10.1128/JVI.75.17.8117-8126.2001

•Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol. 3(9): e300. http://dx.doi.org/10.1371/journal.pbio.0030300

•Jobes DV, Chima SC, Ryschkewitsch CF, Stoner GL (1998). Phylogenetic analysis of 22 complete genomes of the human polyomavirus JC virus. J. Gen. Virol. 79:2491-8.088850.

•Koonin EV, Aravind L, Kondrashov AS (2000). The impact of comparative genomics on our understanding of evolution. Cell. 101(6): 573–576. http://dx.doi.org/10.1016/S0092-8674(00)80867-3

•Luo C, Bueno M, Kant J, Martinson J, Randhawa P (2009) Genotyping schemes for polyomavirus BK, using gene-specific phylogenetic trees and single nucleotide polymorphism analysis. J. Virol. 83(5): 2285–2297. http://dx.doi.org/10.1128/JVI.02180-08

•Maan NS, Mertens PPC, Belaganahalli MN, Singh KP, Nomikou K, Maan S (2013). Full genome sequence of a western reference strain of bluetongue virus serotype 16 from Nigeria. Genome Announc. 1(5):e00684-13.

•Magiorkinis G, Magiorkinis E, Paraskevis D, Vandamme AM, Van Ranst M, Moulton V, Hatzakis A (2004). Phylogenetic analysis of the full-length SARS-CoV sequences: evidence for phylogenetic discordance in three genomic regions. J. Med. Virol. 74(3): 369–372. http://dx.doi.org/10.1002/jmv.20187

•Matthijnssens J, Ciarlet M, Heiman E, Arijs I, Delbeke T, McDonald SM, Palombo EA, Iturriza-Gómara M, Maes P, Patton JT, Rahman M, Van Ranst M (2008). Full genome-based classification of rotaviruses reveals a common origin between human Wa-like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. J. Virol. 82(7): 3204–3219. http://dx.doi.org/10.1128/JVI.02257-07

•Montague MG, Hutchison CA (2000). Gene content phylogeny of herpesviruses. Proc. Natl. Acad. Sci. USA. 97(10): 5334-5339. http://dx.doi.org/10.1073/pnas.97.10.5334

•Olvera A, Cortey M, Segales J (2007). Molecular evolution of porcine circovirus type 2 genomes: phylogeny and clonality. Virology. 357: 175–185.

•Olvera A, Busquets N, Cortey M, de Deus N, Ganges L, Núñez JI, Peralta B, Toskano J, Dolz R (2010). Applying phylogenetic analysis to viral livestock diseases: Moving beyond molecular typing. Vet. J. 184(2): 130–137. http://dx.doi.org/10.1016/j.tvjl.2009.02.015

•Padhi A, Poss M (2009). Population dynamics and rates of molecular evolution of a recently emerged paramyxovirus, avian metapneumovirus subtype C. J. Virol. 83(4): 2015–2019. http://dx.doi.org/10.1128/JVI.02047-08

•Rokas A, Holland PWH (2000). Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15(11): 454–459. http://dx.doi.org/10.1016/S0169-5347(00)01967-4

•Shackelton LA, Holmes EC (2004). The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol. 12(10): 458-465. http://dx.doi.org/10.1016/j.tim.2004.08.005

•Snel B, Huynen MA, Dutilh BE (2005). Genome trees and the nature of genome evolution. Ann. Rev. Microbiol. 59: 191-209. http://dx.doi.org/10.1146/annurev.micro.59.030804.121233

•Sugimoto C, Kitamura T, Guo J, Al-Ahdal MN, Shchelkunov SN, Otova B, Ondrejka P, Chollet JY, El-Safi S, Ettayebi M, Senguet G, Kocagos ZT, Chaiyarasamee S, Thant KZ, Thein S, Moe K, Kobayashi N, Taguchi F, Yogo Y (1997). Typing of urinary JC virus DNA offers a novel means of tracing human migrations. . Proc. Natl. Acad. Sci. USA. 94: 9191-9196. http://dx.doi.org/10.1073/pnas.94.17.9191

•Tatte VS, Rawal KN, Chitambar SD (2010). Sequence and phylogenetic analysis of the VP6 and NSP4 genes of human rotavirus strains: evidence of discordance in their genetic linkage. Infect. Genet. Evol. 10(7): 940–949. http://dx.doi.org/10.1016/j.meegid.2010.05.017

•Thongprachum A, Chan-it W, Khamrin P, Okitsu S, Nishimura S, Kikuta H, Yamamoto A, Sugita K, Baba T, Mizuguchi M, Maneekarn N, Hayakawa S, Ushijima H (2013). Reemergence of new variant G3 rotavirus in Japanese pediatric patients, 2009– 2011. Infect. Genet. Evol. 13: 168–174. http://dx.doi.org/10.1016/j.meegid.2012.09.010

•Yang T, Liu N, Xu Q, Sun E, Qin Y, Zhao J, Wu D (2011). Complete genomic sequence of bluetongue virus serotype 16 from China. J. Virol. 85(24): 13472. http://dx.doi.org/10.1128/JVI.06402-11

•Wang S, Luo X, Wei W, Zheng Y, Dou Y, Xuepeng Cai (2013).

Calculation of evolutionary correlation between individual genes and full-length genome: a method useful for choosing phylogenetic markers for molecular epidemiology. PLoS ONE 8(12): e81106. http://dx.doi.org/10.1371/journal.pone.0081106

• Wang Y-H, Pang B-B, Ghosh S, Zhou X, Shintani T, Urushibara N2, Song YW, He MY3, Liu MQ1, Tang WF1, Peng JS1, Hu Q1, Zhou DJ1, Kobayashi N2. (2014) Molecular Epidemiology and Genetic Evolution of the Whole Genome of G3P [8] Human Rotavirus in Wuhan, China, from 2000 through 2013. PLoS ONE 9(3): e88850. http://dx.doi.org/10.1371/journal.pone.0088850